A NEW APPROACH TO IMPUTATION

Michael P. Cohen, National Center for Education Statistics*
555 New Jersey Avenue NW, Washington, DC 20208-5654

Key Words: Missing data, Covariance, Item nonresponse

Abstract: Imputation is a popular method of handling item nonresponse. With common methods of imputation, though, the usual variance formulae understate the variance of estimates. This paper proposes that items be imputed from distributions more diffuse than those of the real data, thereby compensating for the underestimation of variance by the usual formulae. The impact on covariances is considered in the design of the method. The method is intended for use by data analysts applying techniques based on first and second moments of means only.

1. Introduction

Most surveys have item nonresponse no matter how well planned they may be. These missing data become a problem when it comes time to analyze the dataset. There are three main methods for dealing with item-level missing data: (i) delete complete cases whenever there are missing data for *any* variable being analyzed, (ii) delete cases but only as necessary for a particular family of estimates, and (iii) impute ("fill in values for") the missing data. Methods (i) and (ii) are still widely used in the social and behavioral sciences. Method (iii), though, has been demonstrated to be superior in previous research (Chan and Dunn, 1972; Beale and Little, 1975; Kim and Curry, 1977; Little, 1988; and Bello, 1995).

The problems with methods (i) and (ii) are not hard to ascertain. Method (i) may result in a substantial loss of cases, especially when many variables are being analyzed. The cases retained, moreover, may not be representative of those deleted, resulting in a bias. Method (ii) has the problems of method (i) but to a lesser degree. It has the serious additional problem of inconsistencies in the values of estimates. For instance, if x is being analyzed in conjunction with y, then the estimated mean of x will be based on cases where neither x nor y is missing. If, in another analysis, x is analyzed in conjunction with z instead, the estimated mean of x will in general be based on different cases so we get two different estimates of the same quantity. These inconsistencies can be very confusing to careful readers, resulting in a loss of confidence in the research.

Method (iii), called the *imputation technique*, solves the problems alluded to above. After imputation, one can use complete data methods of analysis without any need to discard cases. Another advantage is that the data can be imputed "in house," thus bringing the additional knowledge of the data collection people to bear on the missing data problem. This is not to say that imputation does not have its own drawbacks. Chief among these is the underestimation of standard errors — this happens essentially because the amount of "real" data is less than it appears to be. Although the reason is less obvious, covariance estimates undergo shrinkage toward zero (i. e., *attenuation*). These matters will be treated in more detail subsequently.

For general discussion of imputation, we recommend Kalton (1983), Kalton and Kasprzyk (1986), and Rubin (1987).

The outline of this paper is as follows: Section 1 is this introduction. In Section 2 we discuss the one-variable case. The section consists of a subsection on the problems with the traditional approach followed by a subsection on the alternative approach. Section 3 expands the coverage to the many-variable case and, in particular, to the difficult problem of covariances. In the last section we make some final remarks.

2. The One Variable Case

2.1 Problems with the Traditional Approach

We begin by assuming the sample has been divided into groups of observations called *imputation classes* (Kalton, 1983, p. 67). Within each imputation class, we assume for now that the responding units for item y are a random subsample of all sampled units. Let the sample size in the imputation class be n with r responding and $m = n - r$ missing. We can number the units so that units $i = 1, 2, \ldots, r$ responded to item y and units $i = r+1, \ldots, n$ did not. The best estimate (in many respects) of the mean of y within the imputation class is

$$\bar{y}_r = \frac{1}{r}\sum_{i=1}^{r} y_i$$

and the best estimate of the variance of the mean is

$$s^2_{\bar{y}_r} = \frac{1}{r(r-1)}\sum_{i=1}^{r}(y_i - \bar{y})^2.$$

For simplicity we are ignoring the sample weights in this discussion, but they could be incorporated. A finite population correction could also be included.

It is tempting to impute the missing values by $\bar{y}_r$. In fact, this choice has good "first order" properties in that $\frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}_r$. On the other hand,

$$\begin{aligned}\frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i-\bar{y}_r)^2 &= \frac{1}{n(n-1)}\sum_{i=1}^{r}(y_i-\bar{y}_r)^2 \\ &= \frac{r(r-1)}{n(n-1)}s^2_{\bar{y}_r},\end{aligned}$$

so the variance of the mean will be underestimated. This perhaps should not be surprising in that we have chosen to impute the value that minimizes the variance expression.

To combat the problem of underestimation of variances, imputation methods have been proposed that attempt to impute values drawn from the distribution of the observed y's. Although an improvement on mean imputation, this approach is also doomed to failure when it comes to estimation of variances as we shall see. If the imputed $y_{r+1},\ldots,y_n$ are distributed like the observed $y_1,\ldots,y_r$, then $\frac{1}{n-r}\sum_{i=r+1}^{n} y_i \stackrel{E}{=} \bar{y}_r$ so that $\bar{y} \stackrel{E}{=} \bar{y}_r$ where $\stackrel{E}{=}$ denotes "equal in expected value" and $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ is the overall sample mean (in the imputation class). So, like mean imputation, the "first order" properties of these methods are good. Furthermore,

$$\frac{1}{r-1}\sum_{i=1}^{r}(y_i-\bar{y}_r)^2 \stackrel{E}{=} \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2.$$

The variance of the mean is estimated by

$$\begin{aligned}\frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i-\bar{y})^2 &\stackrel{E}{=} \frac{1}{n(n-1)}\frac{n-1}{r-1}\sum_{i=1}^{r}(y_i-\bar{y}_r)^2 \\ &= \frac{r}{n}s^2_{\bar{y}_r}.\end{aligned}$$

The variance of the mean is still underestimated although not so badly as with mean imputation. The problem is that the variance formulae are designed for n "real" observations, not $r < n$ observations and $m = n - r$ imputed values.

What, then, can be done? One promising approach is to alter the variance formula used (Rao and Shao, 1992; Särndal, 1992; Fay, 1996b; and Rao, 1996) but impute only once. Another idea, *multiple imputation*, makes use of several imputations to try to capture the missing variance component in variance estimates when missing data are present (Rubin, 1978, 1996; and Fay, 1992). Fay (1996a) and Kaufman (1996) investigate methods that are mixtures of these two approaches. The challenge is to find a method that is reasonably appealing to social science analysts who are inclined to delete cases to avoid the complications caused by missing data.

We consider in this paper single (as opposed to multiple) imputation methods that are intended for use with the standard variance formulae. The imputed values will be *more* dispersed that the observed values. Clearly this method will not work for estimating all features of the distribution; for example, it is not suited for estimating percentiles or histograms. But many statistical procedures depend on only the first two moments of the distribution (e.g. estimating means, totals, and functions thereof), and it is for these procedures the imputation method is intended.

2.2 The Alternative Approach

Let us try to find imputed values $y_{r+1},\ldots,y_n$ so that

$$\bar{y} = \bar{y}_r \tag{2.1}$$

and

$$\frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{r(r-1)}\sum_{i=1}^{r}(y_i-\bar{y})^2. \tag{2.2}$$

Let

$$D_r^2 = \frac{1}{r}\sum_{i=1}^{r}(y_i-\bar{y})^2$$

and

$$D_m^2 = \frac{1}{m}\sum_{i=r+1}^{n}(y_i-\bar{y})^2.$$

Then D_r^2 and D_m^2 are respectively the average squared deviation of the observed and imputed values about their (common) mean. Rewriting (2.2) in terms of D_r^2 and D_m^2, we have

$$\frac{1}{n(n-1)}(rD_r^2 + mD_m^2) = \frac{1}{r-1}D_r^2. \tag{2.3}$$

Simplifying (2.3), we get

$$D_m^2 = \frac{n+r-1}{r-1}D_r^2. \tag{2.4}$$

There are many solutions to (2.1) and (2.4), but, if $m = n - r$ is even, there is one particularly simple solution:

$$y_i = \bar{y} \pm \sqrt{\frac{n+r-1}{r-1}}D_r \quad \text{for } i = r+1,\ldots,n, \tag{2.5}$$

where $m/2$ imputed values have the + sign and $m/2$ have the − sign. Note that if m is small so that $r \approx n$ then (2.5) reduces to $y_i \approx \bar{y} \pm \sqrt{2}D_r$ for $i = r+1, \ldots, n$; that is, the distance of an imputed value from the mean is about $\sqrt{2}$ times the root mean squared deviation of the observed values. If $m \approx r \approx n/2$, representing a large amount of imputation, we have $y_i \approx \bar{y} \pm \sqrt{3}D_r$ for $i = r+1, \ldots, n$; in this case the distance of an imputed value from the mean is about $\sqrt{3}$ times the root mean squared deviation of the observed values. Certainly the imputed values are much more dispersed than the observed values.

A result like (2.5), but with a finite population correction, was obtained by Lanke (1983) and discussed by Sedransk (1985).

It ought to be mentioned that these imputed values will not generally satisfy the edit checks that the observed values had to satisfy, nor even necessarily be feasible values. But if the signs of the deviations from the mean of the imputed values are assigned randomly (or according to an appropriate pattern), the chance an estimated mean over a reasonably large domain will be outside the variables's range will be very small.

3. The Many Variable Case

Of course, in major surveys we almost always have many variables available to use as covariates but themselves having missing values. Let us begin with the simplest such case.

3.1 Two Variables: Same Units with Missing Values

We assume again the sample has been divided into imputation classes. Within the imputation class, suppose the responding units for items x and y are a random subsample of all sampled units. We further suppose in this subsection that x and y are observed for the same units and missing for the same units. Let the sample size in the imputation class be n with r units responding to the two items and $m = n - r$ missing the two items. We number the units so that units $i = 1, 2, \ldots, r$ responded to items x and y whereas units $i = r+1, \ldots, n$ did not.

We seek to impute so that the means of x and y within the imputation class are $\bar{x}_r = \frac{1}{r}\sum_{i=1}^{r} x_i$ and $\bar{y}_r = \frac{1}{r}\sum_{i=1}^{r} y_i$. We also want

$$\begin{aligned} s^2_{\bar{x}} &\equiv \frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{r(r-1)}\sum_{i=1}^{r}(x_i - \bar{x})^2 \text{ and} \end{aligned}$$

$$\begin{aligned} s^2_{\bar{y}} &\equiv \frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &= \frac{1}{r(r-1)}\sum_{i=1}^{r}(y_i - \bar{y})^2. \end{aligned}$$

Lastly, we would like to preserve the correlation of the means:

$$\begin{aligned} \rho_{\bar{x},\bar{y}} &\equiv \frac{1}{n(n-1)}\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_{\bar{x}} s_{\bar{y}}} \\ &= \frac{1}{r(r-1)}\frac{\sum_{i=1}^{r}(x_i - \bar{x})(y_i - \bar{y})}{s_{\bar{x}} s_{\bar{y}}}. \end{aligned}$$

Let

$$\begin{aligned} D^2_{x,r} &= \frac{1}{r}\sum_{i=1}^{r}(x_i - \bar{x})^2, \\ D^2_{y,r} &= \frac{1}{r}\sum_{i=1}^{r}(y_i - \bar{y})^2, \text{ and} \\ C_{x,y,r} &= \frac{1}{r}\sum_{i=1}^{r}(x_i - \bar{x})(y_i - \bar{y}). \end{aligned}$$

We concentrate first on the case $m = 4$, that is, four pairs of missing values. Let $\dot{x}_j = x_{r+j} - \bar{x}$ and $\dot{y}_j = y_{r+j} - \bar{y}$ for $j = 1, 2, 3, 4$ denote the differences of the missing values from the appropriate mean. Then, by the argument used to get (2.4), we have

$$\begin{aligned} \dot{x}_1 + \dot{x}_2 + \dot{x}_3 + \dot{x}_4 &= 0, \\ \dot{y}_1 + \dot{y}_2 + \dot{y}_3 + \dot{y}_4 &= 0, \\ \dot{x}_1^2 + \dot{x}_2^2 + \dot{x}_3^2 + \dot{x}_4^2 &= 4\frac{n+r-1}{r-1}D^2_{x,r}, \\ \dot{y}_1^2 + \dot{y}_2^2 + \dot{y}_3^2 + \dot{y}_4^2 &= 4\frac{n+r-1}{r-1}D^2_{y,r}, \text{ and} \\ \dot{x}_1\dot{y}_1 + \dot{x}_2\dot{y}_2 & \\ \quad + \dot{x}_3\dot{y}_3 + \dot{x}_4\dot{y}_4 &= 4\frac{n+r-1}{r-1}C_{x,y,r}. \end{aligned}$$

To solve, let's try the trigonometric substitutions

$$\begin{aligned} \dot{x}_1 = -\dot{x}_3 &= \sqrt{2\frac{n+r-1}{r-1}}D_{x,r}\sin\theta, \\ \dot{x}_2 = -\dot{x}_4 &= \sqrt{2\frac{n+r-1}{r-1}}D_{x,r}\cos\theta, \\ \dot{y}_1 = -\dot{y}_3 &= \sqrt{2\frac{n+r-1}{r-1}}D_{y,r}\cos\phi, \text{ and} \\ \dot{y}_2 = -\dot{y}_4 &= \sqrt{2\frac{n+r-1}{r-1}}D_{y,r}\sin\phi. \end{aligned}$$

One can verify that all equations are satisfied provided that

$$\begin{aligned} D_{x,r}D_{y,r}(\sin\theta\cos\phi + \cos\theta\sin\phi) & \\ = D_{x,r}D_{y,r}\sin(\theta+\phi) & \\ = C_{x,y,r}. & \end{aligned}$$

So

$$\theta + \phi = \arcsin\left(\frac{C_{x,y,r}}{D_{x,r}D_{y,r}}\right).$$

It is easy to check that the argument of the arcsin function is at most 1 in absolute value so $\theta+\phi$ is well defined. So long as the constraint on their sum is satisfied, θ and ϕ may take on a range of values, each corresponding to a solution to the original equations.

Now let's turn to the harder case (because it is less symmetric): $m = 3$. The equations for this case are

$$\begin{aligned}
\dot{x}_1 + \dot{x}_2 + \dot{x}_3 &= 0, \\
\dot{y}_1 + \dot{y}_2 + \dot{y}_3 &= 0, \\
\dot{x}_1^2 + \dot{x}_2^2 + \dot{x}_3^2 &= 3\frac{n+r-1}{r-1}D_{x,r}^2, \\
\dot{y}_1^2 + \dot{y}_2^2 + \dot{y}_3^2 &= 3\frac{n+r-1}{r-1}D_{y,r}^2, \text{ and} \\
\dot{x}_1\dot{y}_1 + \dot{x}_2\dot{y}_2 + \dot{x}_3\dot{y}_3 &= 3\frac{n+r-1}{r-1}C_{x,y,r}.
\end{aligned}$$

A particular solution, obtained through the use of substitutions and a careful examination of the solution to certain quadratic equations, is given by

$$\begin{aligned}
\dot{x}_1 &= -\sqrt{\frac{n+r-1}{2(r-1)}}D_{x,r}, \\
\dot{x}_2 &= \sqrt{2\frac{n+r-1}{r-1}}D_{x,r}, \\
\dot{x}_3 &= -\dot{x}_1 - \dot{x}_2, \\
\dot{y}_1 &= -\sqrt{\frac{n+r-1}{2(r-1)}}\left(\frac{C_{x,y,r}}{D_{x,r}}\right) \\
&\quad + \sqrt{\frac{3(n+r-1)}{2(r-1)}\left(1 - \frac{C_{x,y,r}^2}{D_{x,r}^2 D_{y,r}^2}\right)}\, D_{y,r}, \\
\dot{y}_2 &= \sqrt{2\frac{n+r-1}{r-1}}\left(\frac{C_{x,y,r}}{D_{x,r}}\right), \text{ and} \\
\dot{y}_3 &= -\dot{y}_1 - \dot{y}_2.
\end{aligned}$$

The cases $m > 4$ are easier and can be handled in a variety of ways. For example, one way to treat $m = 5$, although probably not the best way, is to set $\dot{x}_5 = \dot{y}_5 = 0$ and then apply the solution for $m = 4$.

For $m = 1$ and $m = 2$, no exact solutions can be obtained. If the correlation between $\bar{x}$ and $\bar{y}$ is important, we recommend dealing with $m = 1$ or $m = 2$ by making a random choice among the solutions for $m = 3$ or $m = 4$.

3.2 Two Variables: Only One Variable Missing

Within each imputation class, suppose now that item x is observed for all n units. Item y, on the other hand, is missing for $m \geq 1$ units and observed for the other $r = n - m$ units. We assume the missing y's are missing at random but not necessarily missing *completely at random*; that is, the missingness may depend on the observed x's and y's. The units are numbered so that units $i = 1, 2, \ldots, r$ responded to item y whereas units $i = r+1, \ldots, n$ did not.

This situation introduces an important new feature: It is no longer appropriate to assume that $\bar{y}_r$ is the "best" estimate of the population mean of the y's. We can do better by making use of the x's corresponding to the missing y's.

Consider

$$e_i = \frac{y_i}{x_i}, \qquad i = 1, \ldots, n.$$

We shall explore the assumption that the e_i are a random sample, independent of the x's, within the imputation class. This assumption is reasonable in many circumstances and the reasoning can be extended to other situations.

We can apply the results of Subsection 2.2 to impute the "missing" e_i $(i = r+1, \ldots, n)$ to satisfy:

$$\begin{aligned}
\bar{e} = \bar{e}_r &\equiv \frac{1}{r}\sum_{i=1}^{r} e_i, \text{ and} \\
s_{\bar{e}}^2 = s_{\bar{e}_r}^2 &\equiv \frac{1}{r(r-1)}\sum_{i=1}^{r}(e_i - \bar{e})^2.
\end{aligned}$$

From the imputed e_i, we get imputed y_i by $y_i = x_i e_i$.

3.3 More General Situations

We have discussed but a small subset of the multitude of missing data situations that arise in practice. In this subsection we shall just briefly touch upon three aspects needing more serious investigation.

1. We have only considered imputing one or two variables, but there will almost always be more than that, often hundreds. If there are k variables to be imputed, the number of pairwise correlations to consider is $\binom{k}{2} = k(k-1)/2$. Clearly we will reach a point where the equations for the correlations cannot be solved exactly. At least two ways of treating this problem come to mind.

   (a) The variables can be divided into blocks of variables thought to be closely related. We can then try to control only for the correlations between variables within the same block. The presumption is that this will account for most of the correlation.

(b) As an alternative to trying to control certain correlations exactly, we might only seek to control them on average by randomizing among solutions to the equations for the correlations. A related idea would be to seek approximate solutions that minimize the distance (based on some distance function) to the solutions of the individual equations.

2. Even for two variables, we have only considered the two simplest patterns of missingness for the data: either only one of two variables has missing values, or the two variables have missing values for the same units. The hope, of course, is that we can solve more general problems by an iterative procedure, perhaps first imputing values when one variable is missing but not the other, then the reverse, and finally when both are missing.

3. We have treated imputation within imputation classes, implicitly assuming that the imputation will have good properties for means and variances and correlations of means across imputation classes. If the data for each imputation class are (at least approximately) independent from each other, then the assumption is justified. Otherwise, the results presented here can be extended, but only if we know what the variances and correlations of the means of the observed values across imputation classes *should be.*

4. Final Comment

Deletion of cases still seems to be the most common way that data analysts in the social and behavioral sciences cope with item nonresponse. There is therefore value in searching for techniques for handling missing data that are easy to use yet have desirable statistical properties.

This paper is just a beginning exploration of an approach to imputation that makes use of imputed values distributed more diffusely than the observed data. The approach is not intended for all statistical applications, only those based on the first two moments of means. For many problems we hope it will develop into a reliable technique not requiring multiple imputations or special variance formulae.

Acknowledgments: The author thanks Steven Kaufman for invaluable discussions. He also thanks Graham Kalton for pointing out the relevance of Lanke (1983) and Sedransk (1985) and John Eltinge for his discussion of the paper.

REFERENCES

Beale, E. M. L., and Little, R. J. A. (1975). Missing values in Multivariate analysis, *Journal of the Royal Statistical Society B* **37** 129–146.

Bello, A. L. (1995). Imputation techniques in regression analysis: Looking closely at their implementation, *Computational Statistics and Data Analysis* **20** 45–57.

Chan, L. S., and Dunn, O. J. (1972). The treatment of missing values in discriminant analysis – I. The sampling experiment, *Journal of the American Statistical Association* **67** 473–477.

Fay, R. E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 227–232.

Fay, R. E. (1996a). Alternative paradigms for the analysis of imputed survey data, *Journal of the American Statistical Association* **91** 490–498.

Fay, R. E. (1996b). Rejoinder, *Journal of the American Statistical Association* **91** 517–519.

Kalton, G. (1983). *Compensating for Missing Survey Data.* Ann Arbor: University of Michigan.

Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data, *Survey Methodology* **12** 1–16.

Kaufman, S. (1996). Estimating the variance in the presence of imputation using a residual, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, forthcoming.

Kim, J. O., and Curry, J. (1977). The treatment of missing data in multivariate analysis, *Sociological Methods and Research* **6** 215–240.

Lanke, J. (1983). Hot deck imputation techniques that permit standard methods for assessing precision of estimates, *Statistical Review* **21** 105–110.

Little, R. J. A. (1988). Robust estimation of the mean and covariance matrix from data with missing values, *Applied Statistics* **37** 23–38.

Little, R. J. A., and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

Rao, J. N. K. (1996). On variance estimation with imputed survey data, *Journal of the American Statistical Association* **91** 499–506.

Rao, J. N. K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation, *Biometrika* **79** 811–822.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years, *Journal of the American Statistical As-*

sociation **91** 473–489.

Särndal, C. E. (1992). Methods for estimating the precision of survey estimates when imputation has been used, *Survey Methodology* **18** 241–252.

Sedransk, J. (1985). The objectives and practice of imputation, *Proceedings of the First Annual Research Conference,* Bureau of the Census 445–452.

*This paper is intended to promote the exchange of ideas among researchers and policy makers. The views are those of the author, and no official support by the U.S. Department of Education is intended or should be inferred.